Automated Classification of Starch Granules Using Supervised Pattern Recognition of Morphological Properties

Automated classification of starch granules using supervised pattern recognition of morphological properties

Julie Wilson, Karen Hardy, Richard Allen, Les Copeland, Richard Wrangham and Matthew Collins.

Abstract

Image analysis techniques have been used to investigate the likelihood of being able to classify and assign a probability regarding the plant origin of individual starch granules in a collection of granules. Quantifiable variables were used to characterize the granules, and the assignments and probabilities were calculated objectively. We consider the classification of images containing granules of a single species and of mixed species and the possibility of assigning a class to granules of unknown species in an image of a slide obtained from the dental calculus of chimpanzees.

1. Introduction

Starch is the main store of carbohydrate and energy in plants. It is deposited in semi-crystalline granules in most tissues but is particularly abundant in storage tissues such as seeds, roots and tubers. The granules consist almost entirely of two glucose polymers, amylose and amylopectin, with small amounts of lipids, minerals and phosphorus (Buléon *et al*., 1998; Copeland *et al*., 2009). The complexity of starch biosynthesis results in natural variability in amylose and amylopectin molecules, which is reflected in diversity of granule morphology. This variability, notably in granule size and shape, is associated with differences in functional properties in food processing and nutrition (Peterson and Fulcher, 2001) and the possibility of relating granule morphology to manufacturing process or nutritional qualities is of importance to the food industry.

Starch granule morphology is determined by the action of the biosynthetic enzymes. These occur in multiple forms and both the expression of the underlying genetic diversity and the enzyme activities are influenced by external environmental factors that lead to variability of the starch end product (Morell and Myers, 2005). This variability led Reichert (1913) to suggest the potential of granule morphology as a taxonomic tool.

Starch granule analysis is becoming an increasingly widely used resource in bioarchaeological studies. Based largely on the classification of granules in archaeological samples retrieved from sediments, dental calculus and ancient tools and pots starch granules have been used for determining the diet of ancient human cultures (Hardy *et al*., 2009; Henry and Piperno, 2008; Mercader *et al*., 2008; Piperno *et al*., 2000, 2009), the origins of agriculture, plant domestication and trajectories (Balter, 2007; Iriarte *et al.,* 2004; Piperno and Holst, 1998; Denham et al 2003) and ancient tool use (Barton *et al*. 1998; Loy *et al*., 1992,). Pollen and phytolith analyses are also used, often in combination with starch granule analysis (see, for example, Horrocks, 2005). Holst *et al*.

(2007) found fossil starch granules and phytoliths to be more useful than pollen in discriminating wild from domesticated maize. However, in their investigation to recover dietary information using plant microfossils from dental calculus, Henry and Piperno (2008) found that all of the teeth produced starch granules, some in large numbers, but that their sample contained very few phytoliths.

The identification of starch granules based on morphological characteristics have resulted in new interpretations in particular in relation to plant dispersals, for example in Central and South America (Dickau et al 2007), movement and arrivals of people, for example in Oceania (Horrocks *et al*., 2007), earliest evidence of agriculture (for example, Denham *et al.,* 2003) and early processing of wild plants in the Near East (Piperno *et al*., 2004), Europe (Aranguren *et al.,* 2007) and Africa (van Peer *et al.,* 2003). Whilst, starch granule identification is potentially an exciting way to access information on ancient plants, a better understanding of the origins of starch granule shape is required for this potential to be realised and results need to be placed within the context of a lack of explanation for starch granule shape. Currently, how variations in shape occur, the effects of local geographical and environmental effects on granule shape and the diagenetic effects of ageing are not fully understood or explained.

In archaeological studies, starch granules are often assigned to species, based upon their morphological characteristics, but such studies usually do not use formal statistics and instead rely upon the visual comparison between individual archaeological granules with reference granules. However there are a number of potential problems with this approach. In most domestic cereal plant species (for example, corn, wheat, rice), there has been strong selective pressure through plant breeding and selection over thousands of years to enhance starch production. Starch granule phenotype will reflect expanding environmental conditions acting upon an increasingly domesticated genotype, and it is possible that different combinations of factors may result in similar morphological end points.

We believe it would therefore be useful to document the extent of morphological variation within and between species. There has only been one such study directed towards archaeological research; Torrence *et al.* (2004) used measurements obtained interactively from images of starch granules and multivariate analysis for classification. The discriminatory features also include categorical variables requiring subjective decisions by the researcher that may well be influenced by prior knowledge regarding the actual class of a granule.

In the present study we investigate a completely objective classification of images of starch granules from nine genetically diverse species.

2. Materials and Methods

2.1. Starch preparation

Specimens of modern lotus root (*Nelumbo nucifera*) (Gaertn.), maize (*Zea mays* L. ssp. mays), mung bean (*Vigna radiate* (L.) R. Wilczek), oat (*Avena sativa*), plantain

(*Musa.Sp)*, sweet potato (*Ipomoea batata*), potato (*Solanum tuberosum* L.), cassava (*Manihot esculenta*), and water chestnut (*Eleocharis dulcis*) starch granules were prepared for image analysis. Of the samples used, lotus, maize, mung bean, potato, cassava, and water chestnut were highly refined starch powders. Oat, plantain, and sweet potato were prepared from meal (oats) and flesh preserved in alcohol (plantain and sweet potato).

2.2 Image acquisition

For each of the starch types selected for analysis, a set of five slides were mounted with glycerol and left for a minimum of 24 hours at 21(+/- 3)ºC before image analysis. In addition, a set of five mixed starch slides were prepared consisting of plantain and sweet potato. JPEG images at 300 dpi and 24 bit colour depth of the starch grains were obtained using the Digital Imaging Solutions program CellD, version 2.6 (Build 1200) and an Olympus IX71 inverted microscope with fitted ColourView III camera. The magnification was set at x100 (x10 objective).

For each slide a pair of images, consisting of one photograph taken in white light and a corresponding photograph in polarised light, were acquired from two different fields of view. Thus, 10 image pairs were obtained for each set of five slides.

In addition a pair of images were acquired from a sample of dental calculus from recently deceased chimpanzees from the Kibale Chimpanzee Project, Uganda (Carter *et al*., 2008).

2.3. Image processing

The edges of objects in an image give rise to sudden changes in intensity from one pixel to the next and can therefore be identified by analyzing the gradient, or rate of change, of the image intensity. Mathematically the gradient is calculated by differentiation but simple operators can be used to approximate the gradient of an image (see Gonzalez and Woods , 2002, for example). Here we use the Sobel operator to approximate the rate of change at each pixel from its immediate neighbours. Figure 2 shows how connected sets of pixels with gradient magnitudes above a threshold, calculated from the statistics of the image intensities, allows the granules to be identified.

Groups of touching or overlapping granules will form single objects and must be separated. The edges of individual granules can be detected using the zero crossings of the image, which is first filtered with the Laplacian of Gaussian filter. This allows closed contours of single pixel width to be detected and shared boundaries identified as those that cross the object from one external boundary point to another (see Figure 3). Granules that were partially obscured by others lead to incomplete granules and some broken or cracked granules result in erroneous segmentation. However, such objects can be recognised using the shape descriptors described in the next section and eliminated from the analysis. Any granules in contact with the edge of the image are also removed to restrict the analysis to whole granules. Finally, a mask separating the individual granules

from the background is applied to the corresponding polarized image (Figure 4). Features to be used for classification are extracted from each granule using both images. For brevity, in the following sections, we refer to each image pair, the photograph taken in white light and the corresponding photograph taken in polarized light, as the image rather than the pair of images.

The analysis was carried out using software written in-house in C and, with the exception of the elimination of background debris from one image, required no human intervention for starch granule classification.

2.4. Feature extraction

In order to classify the granules, characteristic features that can be quantified must be determined. Many variables have been used in the identification of granules to plant genus or species (e.g. Torrence & Barton, 2006). Torrence *et al*. (2004) define variables related to the size and shape of the granule and the polarization cross as well as a number of features described simply as present or absent. Here, the number of pixels the granule covers determines its area. The greatest distance between boundary pixels is taken as the length of the granule and the maximal distance orthogonal to this as its width. The minimum rectangular box enclosing the granule is determined and the ratio of its area to the area of the granule used as a simple shape descriptor. The ratio of length to width gives further information on granule shape as does the ratio of area to the square of the number of boundary pixels. Here the length of the boundary is squared to remove the effect of granule size on this variable. The variance in distance from the centroid of the granule to its boundary provides another variable and granule shape is also compared to an ellipse with major and minor axes equal to the length and width of the granule. The elliptic variance is defined as the mean-squared error of the granule with respect to the ellipse.

Chain codes (Freeman, 1974) allow the boundaries of objects to be represented in a compact way. The chain code is an ordered sequence of integers encoding the direction of the vector connecting neighbouring boundary pixels. Straight sections of the boundary can be recognized by repeats in the sequence. We use the sum of the squared lengths of all straight segments to give greater weight to longer sections and the length of the longest straight section provides another variable. Chain codes can also be use to calculate curvature. However, as they use the change from one pixel to the next, chain codes are very sensitive to small changes in direction. Figure 5 shows how an estimate of the curvature at any boundary pixel can be obtained by considering two points, one n boundary pixels before it and one n boundary pixels after it. The difference between the straight-line distance between the two points and that along the boundary gives the curvature. The greater the value of n, the less localized the curvature. Here we take $n = 2$ and the average over all pixels taken as a measure of curvature. Two further measures related to curvature are the total concavity, being the difference between the area of the

granule and that of its convex hull and the maximum concavity defined as the greatest distance between the granule boundary and the convex hull as shown n Figure 6.

Torrence *et al.* (2004) define several variables related to the polarization cross in their classification of starch granules. These include categorical variables such as the style of the cross (straight, wavy, figure of eight) and quantitative variables (area of polarization cross, maximum dimension, distance between long arms and angle of the cross) measured using interactive graphics software. Here we estimate the area of the polarization cross by the percentage of pixels below a threshold of $\mu + 0.1\sigma$, where μ and σ are the mean and standard deviation of the pixels in each masked granule (Figure 7). Obtaining other measurements automatically is more complicated. A thinning algorithm was applied to the pixels identified by the cross area to skeletonize the cross. For clear examples such as those shown in Figure 7, this allows the length and curvature of each cross arm and the angles between them to be calculated. However a number of problems arise in the skeletonization of the cross. Cracked granules and those with strong indentations produce extra branches that make further calculations impossible. The presence of multiple disperse branch points could be used to filter out such granules but extra loops and branches also occur for other perfectly good granules as can be seen in Figure 8. Some loops can be recognized and eliminated but others cannot and in these cases all variables relying on the skeletonization must be given null values.

The full list of variables calculated is shown in Table 1. All variables were mean centred and scaled to unit variance (using the mean and standard deviations calculated from the training data) to give equal weight to all variables. The results obtained without using variables that depend on the skeletonisation of the polarization cross (indicated by an asterisk in Table 1) are better than those obtained using all variables. This can be explained by the fact that there are granules of all species for which the skeletonization is deemed unreliable. Therefore the classification results described in the next section were obtained without the last three variables in Table 1.

3. Results

3.1. Granule classification

The large number of observations and limited number of variables make the data suitable for most supervised learning algorithms, which are trained to associate a certain output (the class) with a particular input (the feature vector of classification variables). The way in which the algorithm learns the rules of association differs between classifiers but all supervised methods are prone to over-fitting, i.e. the rules may fit the data used during training, but will not generalise to new observations. The use of a completely independent test set, consisting of data not used during training is therefore vital. Of the 6120 individual granules identified in the images, 100 from each of the nine classes were chosen randomly to form a training set and the remaining 5220 granules used as an independent test set. The extracted feature vectors for the training set were used to train a Support Vector Machine (SVM) classifier with a radial basis function (RBF) kernel.

Other classifiers, including a linear SVM, self-organising maps (SOM) and learning vector quantization (LVQ), were also tried and gave very similar results. Furthermore, combining classifiers using a majority vote system did not improve the results. The confusion matrix in Table 2 shows the classification rates obtained using the trained SVM-RBF classifier on the test set (never used during training). The rows in the table correspond to the actual type of granule in the test set with the number of granules of each type in brackets. The columns indicate the class that the granules were assigned to by the classifier. The figures given are percentages, so for example, 57% of the 548 lotus root granules in the test set were classified correctly, 17% were assigned to class 4 (potato), 1% were assigned to each of class 3 (maize) and class 5 (water chestnut), 2% were assigned to class 6 (cassava), 1% were assigned to each of class 7 (sweet potato) and class 8 (oats) and 9% were classified as class 9 (plantain). Thus the diagonal entries (in bold italic) show the percentages of correct classifications. It can be seen that plantain and potato granules classify reasonably well with 79% and 70% correct, respectively. Although 57% of lotus root granules are classified correctly, many (17%) are confused with potato granules. Most miss-classified mung bean granules are classed as maize (11%) whereas 64% are correctly classified. Similarly, maize granules (45% correct) are confused with mung bean (13%) and also cassava (16%). The classification of the other types of granules, water chestnut, cassava, sweet potato and oats is poor with much confusion between these groups.

Relating the classification rates to the example images in Figure 1, it could be argued that the granules that the best classification rates are obtained for species with granules that tend to be larger. This might suggest that the resolution of the images is not sufficient for discrimination of the smaller granules. However, these groups have the greatest variability of size as can be seen from the statistics in Table 3 and the box-plots in Figure 9. The population pyramid in Figure 10 does show that more than half the smallest granules are miss-classified and that bigger granules are correctly classified more often than not. However, there are miss-classifications for all sizes of granule, including the largest. Moreover, mung bean granules, which classify well, are generally of a similar size to cassava granules, which do not.

3.2. Image classification

In general the classification of individual granules is not very successful. Even with just nine species, there is a great deal of overlap between classes. We now consider classification of a group of granules. A number of granules can be identified in each image and therefore we can obtain a class for each image by combining the classification of the individual granules within that image. Any granules in an image that were used during training are not used here to classify the image. Different numbers of granules, ranging from just five up to 143, are obtained for the images.

3.2.1. Single species

For each single species image, a class was obtained by combining the predicted classes of the individual test set granules in the image. The results shown in Table 4 were obtained using four different classifiers (SVM-RBF, SVM-linear, SOM and LVQ). For every test set granule in an image, each class was assigned a probability of 0.25 for each classifier that assigned the class to the granule. Thus, if any granule was assigned the same class by all four classifiers, that class was given a probability of 1.0 for the granule. The probabilities for each class were added over all test set granules in the image and the class with the highest probability assigned to the image. In this case there was some improvement in the results when classifiers were combined, with four less images classified correctly when a single classifier was used.

The single species with the highest score was taken as the class of the image however close other scores might be. Most miss-classified images had similar scores for different species showing the classification to be unconvincing whereas, for some of the correctly classified images, the highest score identifying the species was significantly higher than other class scores. However, for other correctly classified images, the highest score was not much greater than that of other classes. It would not be clear that such images showed granules of a single species without prior knowledge. The classification of images with mixed species is considered in the next section.

3.2.2. Mixed species

In addition to the single species images, mixed starch images were obtained from slides of both plantain (class 7) and sweet potato (class 9). It can be seen from Table 2 that plantain is one of the easiest species to classify but that sweet potato granules are often miss-classified.

In order to get a true indication of the classification results possible when multiple granules of any type are available, the likelihood of a class being incorrect should be taken into account. Reading down the columns in Table 2, we can see how many of the granules assigned to any class really do belong to that class. Thus 57% of the 548 lotus root granules, i.e. 312 of all granules classified as lotus root, really are lotus root but, for example, 1% of the 777 maize granules are also classified wrongly as lotus root. We can use this information to obtain a probability for each assigned class being correct. For example, the probability that any granule classified as lotus root really is lotus root is calculated as:

$$\frac{57 \times 548}{57 \times 548 + 1 \times 777 + 7 \times 933 + 5 \times 703 + 5 \times 758 + 1 \times 562 + 1 \times 318 + 1 \times 204 + 3 \times 317} \cong 0.66.$$

The probabilities obtained from the test set results are shown in the final row of Table 2.

However, using probabilities obtained from the test set results to assign classes to the images would mean that the results obtained were no longer independent. We therefore use the probabilities obtained from classification of the granules in the training set. The probability of granules of each species being classified correctly by each of the four classifiers is given in Table 5. The differences between classifiers give an indication

as to why combining classifiers can improve the results. Although the SVM-RBF classifier looks like the best single classifier, the differences between the probabilities obtained from the training data (Table 5) and those obtained from the test set (Table 2) show that this classifier has a tendency to overfit the training data. For example, granules classed as oats (class 8) in particular were more likely to be correctly classified during training.

As none of the granules in these images were used to train the classifiers, all may be used to provide the image score. For each image, class scores were obtained for each classifier by multiplying the number of granules in the image assigned to a particular class by the corresponding probability of the class being correct for that classifier. The results for the four classifiers were added to give class scores for the image, which were rescaled so that they summed to 1.0 to represent a probability distribution. Table 6 shows the classification results for the mixed-species images. For each one of the ten images, the probability calculated for each class is shown. It can be seen that class 7 (sweet potato) is given the highest probability for 7 of the 10 images and is second to class 9 (plantain) for another. Plantain appears in the top three most likely species for 7 of the images. Other classes given a high probability for several images are class 6 (cassava) and class 8 (oats). Both of these species have small granules often confused with sweet potato. From these results images 5 and 6 do not look like sweet potato or plantain, although, according to the single species results, plantain granules usually classify well. However neither of these images shows granules that are obviously plantain on inspection by eye and it could be that the area of the slide shown contains only sweet potato granules.

3.2.3. Unknown species

Figure 11(a) shows that the slide of unidentified starch granules from the dental calculus of chimpanzees has a great deal of background debris. The granules are difficult to see by eye in the white light image and it was not possible to identify the granules without further processing. The granules are more obvious in the polarized image (Figure 11(b)) although the boundaries cannot be identified. We therefore combined the two images in order to accentuate the granules before using an interactive graphics programme (GraphicConverter) to trace the granule boundaries and delete the background between them. The mask obtained in this way was then applied to both the white light and polarized images and the processing continued as with all other images from this point. In all, feature vectors from 91 granules were extracted from the image, which were then used to provide the probability of each species being present in the image as described for the mixed-species images. The resulting probabilities are given in Table 7. It can be seen that class 9, plantain, is by far the most likely species (of those in the training set).

3.3 Wider context

Any pattern recognition problem requires features to be extracted from the input data that can be used in some pattern-matching or decision-making procedure for classification. These features should be measurable characteristics that are common within classes and discriminatory between classes. We have found there is considerable variation within species. There are many aspects of the biosynthesis and assembly of granules that are not well understood, and there are other factors that could be important in determining the size and shape of granules (Torrence and Barton, 2006). These include: the intracellular space available for granule deposition; the availability of carbon in excess of immediate metabolic requirements for starch synthesis, that is, whether synthesis is continuous over an extended period or is sporadic or interrupted by periods of degradation; diurnal fluctuations in synthesis; and the optimum size of macromolecules for stability and efficient packaging. Similar biological functionality in Nature may be achieved in different ways, and it is not necessarily surprising that there are similarities among starch granules from the species chosen in this study, despite the considerable genetic diversity between them.

3.3.2 Importance for Archaeological Research

The analysis of starch granules from ancient plants is a rapidly expanding area of archaeological research, but classification often relies on the comparison of individual granules with reference granules. It is therefore important to recognize the shortcomings of this approach. Although some granules of certain species do have distinctive qualities the comparison of discrete granules is likely to be very unreliable. The style of the birefringent cross, clearly visible in polarized light, has been used in other analyses to classify starch granules. In our completely automated analysis, we have not been able to fully exploit the characteristics of the cross. However, Torrence *et al*. (2004) state that their study "may have placed too much emphasis on the characters visible under polarized lighting" and their analysis gave similar results. They considered acentric and centric views separately and obtained overall classification rates of 57% and 75%, respectively. Whilst they were able to classify two species with 100% accuracy for centric views, others had classification rates as low as 14%. Furthermore, the results reported by Torrence *et al*. (2004) are for all data, i.e. including the granules that were used during training rather than for an independent test set. It is not clear how much this has enhanced their results, as the number of granules used during training is not given. The assignment of probabilities to a collection of granules has been shown to be more dependable with some species being particularly consistent. We found that the choice of discriminatory variables is more important than the classifier chosen and a combination of classifiers allows more accurate assignment of probabilities when classifying collections of granules. Although reasonable results were obtained for most images showing granules of a single species, the confusion between species is still evident. Water chestnut granules in particular are frequently miss-classified so that only 4 out of 10 of these images are assigned the correct class using probabilities. For the images of sweet potato and plantain granules together, these two species appear among the top three most

likely for 7 of the 10 images and for another image, although plantain seems unlikely, sweet potato is given the highest probability. However, for two of the mixed plant images, plantain and sweet potato are both given low probabilities. Using the same method, the previously unidentified starch granules from the dental calculus of chimpanzees were given a much higher probability of being plantain than any of the other eight species considered here. It is known that Kibale chimpanzees eat young stems of Musa from fields adjacent to Kibale National Park, and also occasionally eat ripe Musa fruits so the results seem reasonable. Of course, only species for which the classifier has been trained are possible candidates. In fact all classification is a process of elimination and we have only shown that the granules are more likely to be plantain than any of the other species tested.

4. Conclusions

The representation of three-dimensional starch granules in a two-dimensional image has obvious limitations and involves granules in different orientations. This additional variation could affect discrimination, but the use of sufficient training data allows multiple orientations to be represented in machine learning algorithms.

It has been recognised that a 'population' approach to analysis is likely to be more reliable than single granule identification, as any sample of plant starch contains a variety of granule shapes (see Torrence and Barton, 2006). The results here using multiple granules also suggest that, when a number of granules are available, it may be possible to identify certain plants with a degree of security. However, the recovery of multiple granules may not always be possible, and even then not all the granules will necessarily derive from the same plant source. The classification rates obtained here and by others (Torrence *et al*., 2004) show that even with adequate reference collections morphological methods need to be used with caution and that other complementary information would be needed for credible identification of most plant species. Furthermore we suspect that the classification of starch granules from archaeological contexts requires a better understanding of the impacts of both domestication (selection pressure) and diagenesis on granule shape.

5. References

Aranguren B.,Becattini R., Mariotti Lippi M. and Revedin A. (2007). Grinding flour in Upper Palaeolithic Europe (25000 years bp). *Antiquity*, **81**(314):845–855
Balter, M. (2007). Seeking Agriculture's Ancient Roots. *Science*, **29**, 1830-1835.
Barton, H. Torrence, R. and Fullagar, R. (1998). Clues to stone tool function re-examined: comparing starch grain frequencies on used and unused obsidian artefacts. *Journal of Archaeological Science*, **25**, 1231-1238.
Buléon, A. Colonna, P., Blanchot, V. and Ball, S. (1998). Starch granules: structure and biosynthesis. *Int. J. Biol. Macromol*., **23**, 85-112.
Carter, M., Pontzer, H., Wrangham, R. and Kerbis Peterhans, J. (2008). Skeletal

pathology in *Pan troglodytes schweinfurthiin* Kibale National Park, Uganda. *Am. J. Phys. Anthr.* **135**, 431-437.
Copeland, L., Blazek, J., Salman, H. and Tang, M. (2009). Form and functionality of starch. *Food Hydrocolloids,* **23**(6), 1527-1534.
Freeman, H. (1974). Computer processing of line-drawing images. *Computing Surveys*, **6** (1), 57-97.
Denham, T.P., Haberle, S.G., Lentfer, C., Fullagar, R., Field, J., Therin, M., Porch, N. and Winsborough B. (2003). Origins of Agriculture at Kuk Swamp in the Highlands of New Guinea. *Science,* **301**, 189-193.
Dickau, R., Ranere, A.J. and Cooke, R.G. (2007). Starch grain evidence for the preceramic dispersals of maize and root crops into tropical dry and humid forests of Panama. *Proc. Nat. Acad. Sci*. **104**(9): 3651–3656
Gonzalez, R. and Woods, R. (2002). *Digital Image Processing*. Second Edition. Prentice Hall.
Hardy, K., Blakeney, T., Copeland, L., Kirkham, J. Wrangham, R. and Collins, M. (2009). Starch granules, dental calculus and new perspectives on ancient diet. *Journal of Archaeological Science*, **36**, 248-255.
Henry, A. and Piperno, D. (2008). Using plant fossils from dental calculus to recover human diet: a case study from Tell al-Raqã'i, Syria. *Journal of Archaeological Science*, **35**, 1943-1950.
Holst, I., Moreno, J. and Piperno, D. (2007). Identification of teosinte, maize and Tripsacum in mesoamerica by using pollen, starch grains and phytoliths. *Proc. Natl. Acad. Sci*. USA, **104**, 17608-17613.
Horrocks, M. (2005). A combined procedure for recovering phytoliths and starch residues from soils, sedimentary deposits and similar materials *Journal of Archaeological Science* 32.
Horrocks, M., Grant-Mackie, J. Matisoo-Smith, E. (2008). Introduced taro (Colocasia esculenta) and yams (Dioscorea spp.) in Podtanean (2700-1800 years BP) deposits from Mé Auré Cave (WMD007), Moindou, New Caledonia *Journal of Archaeological Science,* **35**, 169-180.
Iriarte, J., Holst I., Marozzi, O., Listopad, C., Alonso, E., Rinderknecht, A. and Montaña, J. (2004). Evidence for cultivar adoption and emerging complexity during the mid-Holocene in the La Plata basin. *Nature,* **432**, 614-617.
Loy T. H., Spriggs M. and Wickler S. (1992). Direct evidence for human use of plants 28,000 years ago: starch residues on stone artefacts from the northern Solomon Islands. *Antiquity*, **66**(253), 898–912.
Mercader,, J., Bennett, T. and Raja, M. (2008). Middle stone age starch acquisition in the Niassa Rift, Mozambique. *Quaternary Research*, **70**, 283-300.
Morell, M. and Myers, A. (2005). Towards the rational design of cereal starches. *Curr. Opin. Plant Biol*., **8**, 204-210.
Peterson, D. and Fulcher, R. (2001). Variation in Minnesota HRS wheats: starch granule size distribution. *Food Research International*, **34**, 357-363.

Piperno, D. and Holst, I. (1998). The presence of starch grains on prehistoric stone tools from the humid neotropics: indications of early tuber use and agriculture in Panama.
Piperno, D., Ranere, A., Holst, I. and Hansell P. (2000). Starch grains reveal early root crop horticulture in the Panamanian tropical forest.on prehistoric stone tools from the humid neotropics: indications of early tuber use and agriculture in Panama. *Nature*, **407**, 894-897.
Piperno, D., Ranere A., Holst, I., Iriarte, J. and Dickau, R. (2009). Starch grain and phytolith evidence for early ninth millennium B.P. maize from the Central Balsas River Valley, Mexico. *Proc. Natl. Acad. Sci*. USA, **106**, 5019 –5024.
Journal of Archaeological Science, **25**, 765-776.
Reichert, E. (1913). *The differentiation and specificity of starches in relation to genera, species, etc.* Washington. Carnegie Institute.
Torrence, R., Wright, R. and Conway, R. (2004). Identification of starch granules using image analysis and multivariate techniques. *Journal of Archaeological Science*, **31**, 519-532.
Torrence R & Barton H. (2006). *Ancient Starch Research*, Left Coast Press.
Van Peer, P., Fullagar, R., Stokes, S., Bailey, R.M., Moeyersons, J., Steenhoudt, F.,Geerts, A., Vanderbeken, T., De Dapper, M. and Geus, F. (2003). The Early to Middle Stone Age transition and the emergence of modern human behaviour at site 8- B-11. Sai Island, Sudan. *J. Hum. Evol*., **45**, 187–193.

Table 1: Summary of variables used for classification. Variables marked with an asterisk rely on the skeletonization of the polarization cross and were not used to obtain the results described in section 5.

size	granule area
	granule length
	granule width
shape	granule length to width ratio
	minimum box area to granule area ratio
	squared boundary length to area ratio
	variance in distance from boundary to centroid of granule
	elliptic variance
curvature	overall straightness of boundary
	length of maximum straight section as fraction of boundary
	total curvature
	total concavity
	maximum concavity
polarization cross	area of cross to granule area ratio
	total curvature of cross arms*
	distance from cross centre to centroid of granule*
	minimum angle between cross arms*

Table 2: Classification rates for starch granules in the test set. Diagonal elements in bold italic show the percentages of correctly classified granules.

species (#)	actual class	assigned class								
		lotus	maize	mung	potato	water chestnut	cassava	sweet potato	oats	plantain
		1	2	3	4	5	6	7	8	9
lotus (548)	1	***57***	0	6	17	6	2	1	1	9
maize (777)	2	1	***45***	13	1	6	16	9	9	3
mung bean (933)	3	7	11	***64***	3	6	4	2	2	2
potato (703)	4	5	0	7	***70***	7	0	3	3	5
water chestnut (758)	5	5	16	11	4	***22***	14	13	11	2
cassava (562)	6	2	17	6	2	10	***32***	17	6	10
sweet potato (318)	7	1	11	2	3	12	12	***35***	17	7
oats (204)	8	1	12	6	10	11	4	21	***30***	5
plantain (317)	9	3	3	1	3	1	3	3	5	**79**
probability assigned class is correct		0.65	0.47	0.65	0.70	0.35	0.35	0.24	0.17	0.52

Table 3: Mean and standard deviation of granule size (as determined by granule length in pixels) for each species.

species	mean size	standard deviation of size
lotus	807.5	515.4
maize	249.9	97.9
mung bean	363.4	169.1
potato	690.9	838.1
water chestnut	296.3	149.3
cassava	306.4	143.8
sweet potato	274.4	150.7
oats	242.7	126.9
plantain	611.2	336.2

Table 4: Classification of starch images obtained from the assigned classes of the test set granules in each image. Figures given are the numbers of images where there are 10 images in total for each species.

Matthew Collins 8/8/11 11:33 AM
I have added the species name underneath

species		assigned class								
		lotus	maize	mung	potato	water chestnut	cassava	sweet potato	oats	plantain
Actual class		1	2	3	4	5	6	7	8	9
lotus	1	***10***	0	0	0	0	0	0	0	0
maize	2	0	***8***	0	0	0	2	0	0	0
mung bean	3	0	0	***10***	0	0	0	0	0	0
potato	4	0	0	0	***10***	0	0	0	0	0
water chestnut	5	0	1	2	0	***4***	3	0	0	0
cassava	6	0	1	1	0	1	***7***	0	0	0
sweet potato	7	0	1	0	0	3	0	***6***	0	0
oats	8	0	0	0	1	1	0	2	***6***	0
plantain	9	3	0	0	0	0	0	0	0	**10**

Table 5: Conditional probabilities of the class assigned by each classifier being correct. The

	assigned class								
	1	2	3	4	5	6	7	8	9
SVM-RBF	0.85	0.54	0.74	0.73	0.59	0.53	0.48	0.64	0.82
SVM-Linear	0.62	0.42	0.61	0.68	0.32	0.29	0.25	0.13	0.51
SOM	0.46	0.40	0.52	0.58	0.27	0.27	0.21	0.21	0.47
LVQ	0.50	0.45	0.47	0.57	0.27	0.25	0.16	0.14	0.37

probabilities were obtained from classification of the training set granules.

image	class (probability)								
1	**7** (0.41)	8 (0.26)	6 (0.12)	2 (0.11)	**9** (0.04)	5 (0.04)	1 (0.01)	3 (0.00)	4 (0.00)
2	**9** (0.35)	**7** (0.25)	8 (0.17)	2 (0.08)	6 (0.07)	1 (0.04)	4 (0.02)	5 (0.02)	3 (0.00)
3	**7** (0.40)	**9** (0.15)	6 (0.15)	8 (0.15)	2 (0.09)	5 (0.03)	1 (0.02)	4 (0.01)	3 (0.01)
4	**7** (0.30)	8 (0.23)	**9** (0.22)	6 (0.15)	2 (0.03)	5 (0.02)	3 (0.02)	1 (0.01)	4 (0.00)
5	3 (0.24)	5 (0.23)	2 (0.18)	8 (0.13)	6 (0.11)	**7** (0.09)	**9** (0.02)	4 (0.00)	1 (0.00)
6	5 (0.17)	6 (0.17)	2 (0.16)	**9** (0.15)	3 (0.13)	**7** (0.09)	8 (0.06)	1 (0.05)	4 (0.03)
7	**7** (0.29)	6 (0.28)	**9** (0.11)	8 (0.11)	2 (0.09)	5 (0.04)	3 (0.04)	4 (0.03)	1 (0.02)
8	**7** (0.35)	**9** (0.25)	8 (0.16)	6 (0.12)	2 (0.05)	5 (0.04)	1 (0.01)	3 (0.01)	4 (0.01)
9	**7** (0.26)	**9** (0.24)	6 (0.16)	4 (0.08)	2 (0.08)	8 (0.07)	5 (0.05)	1 (0.05)	3 (0.02)
10	**7** (0.33)	8 (0.16)	**9** (0.15)	6 (0.13)	2 (0.09)	4 (0.07)	5 (0.05)	3 (0.01)	1 (0.01)

Table 6: Results of mixed species analysis: the classes are ordered according to the likelihood (as determined by the class probability in brackets) that the image consists of granules of the corresponding species. These ten images show granules of both plantain and sweet potato. Class 7 corresponds to sweet potato and class 9 to plantain, both class numbers are shown in bold type in the table.

class number	species	probability
9	**plantain**	**0.56**
6	cassava	0.14
1	lotus root	0.11
7	**sweet potato**	**0.07**
8	oats	0.05
2	maize	0.03
3	mung bean	0.01
5	water chestnut	0.01
4	potato	0.01

Table 7: Probabilities for the presences of granules of each species in the slide prepared from the dental calculus of chimpanzees.

Figure Legends

Figure 1: Example images from each of the 9 types of starch granule used in this study: (a) lotus root, (b) maize, (c) mung bean, (d) potato, (e) water chestnut, (f) cassova, (g) sweet potato, (h) oats and (i) plantain. Each image section shown here is approximately one quarter of the original image.

Figure 2: The gradient magnitudes for the image in (a) are shown in (b). The pixels with magnitudes above the threshold are shown in black in (c) and the objects obtained after filling in holes are shown in (d).

Figure 3: The gradient magnitudes obtained for the image section in (a) are shown in (b). Identification of the shared boundaries from the zero crossings in (c) allows the granules to be separated as shown in (d).

Figure 4: Masked images showing individual potato granules. The white light image is shown in (a) and the corresponding polarized image in (b).

Figure 5: A measure of curvature at the pixel indicated by the grey point can be obtained from the difference in length between the sum of the two dotted lines and the dashed line. In (a) this difference is $\sqrt{5}+2-\sqrt{17}=0.113$, whereas in (b) we have $4\sqrt{2}-4=1.657$, reflecting the greater curvature.

Figure 6: The black line shows the convex hull of the granule, i.e. the smallest convex polygon that contains all the pixels of the granule. The total concavity is defined as the difference between the area of a granule and its convex hull. The maximum concavity is defined as the greatest distance between the boundary of the granule and its convex hull, as indicated by the arrow.

Figure 7: A section of an original polarized image is shown in (a) with the corresponding section shown in (b) after the mask obtained from the white light image has been applied. In (c) the black pixels are those below the threshold, $\mu + 0.1\sigma$, where μ and σ are the mean and standard deviation of the pixels in each masked granule. The percentage of pixels that are black pixels provides a classification variable for each granule. The skeletonized image of the polarization cross is shown in (d).

Figure 8: The cracked granules seen in the white light image in (a) and corresponding polarized image in (b) give rise to extra branches in the skeletonized polarisation cross as shown in (c). Extra loops and branches also result from the thresholding procedure. The cross areas obtained for the granules in the polarized image in (d) are shown in (e) with the resulting skeletonization in (f).

Figure 9: Box-plots showing the distribution of granules sizes for each of the 9 species of granule. Granule sizes are given by area as a number of pixels where 200μm corresponds to ~390 pixels.

Figure 10: Population pyramid showing the distribution of correctly classified (type 1) and incorrectly classified (type 0) granules according to size. Granule sizes are given by area as a number of pixels where 200μm corresponds to ~390 pixels.

Figure 11. The white light and polarized images for the chimp slide are shown in (a) and (b) respectively. The image obtained by combining these two images in shown in (c) with the

resulting masked granules in (d). Each image section shown here is approximately one quarter of the original image. The full image is shown in (e) to give the scale of the granules.

Figure 1: Example images from each of the 9 types of starch granule used in this study: (a) lotus root, (b) maize, (c) mung bean, (d) potato, (e) water chestnut, (f) cassova, (g) sweet potato, (h) oats and (i) plantain. Each image section shown here is approximately one quarter of the original image.

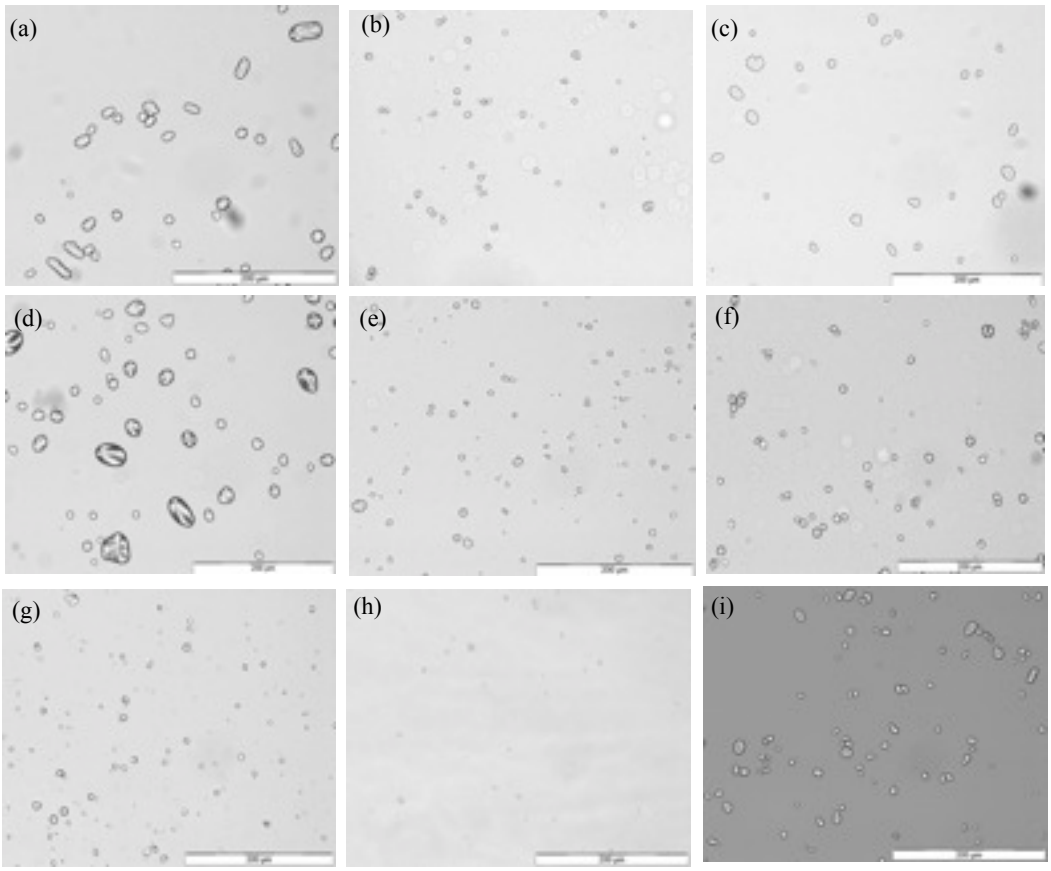

Figure 2: The gradient magnitudes for the image in (a) are shown in (b). The pixels with magnitudes above the threshold are shown in black in (c) and the objects obtained after filling in holes are shown in (d).

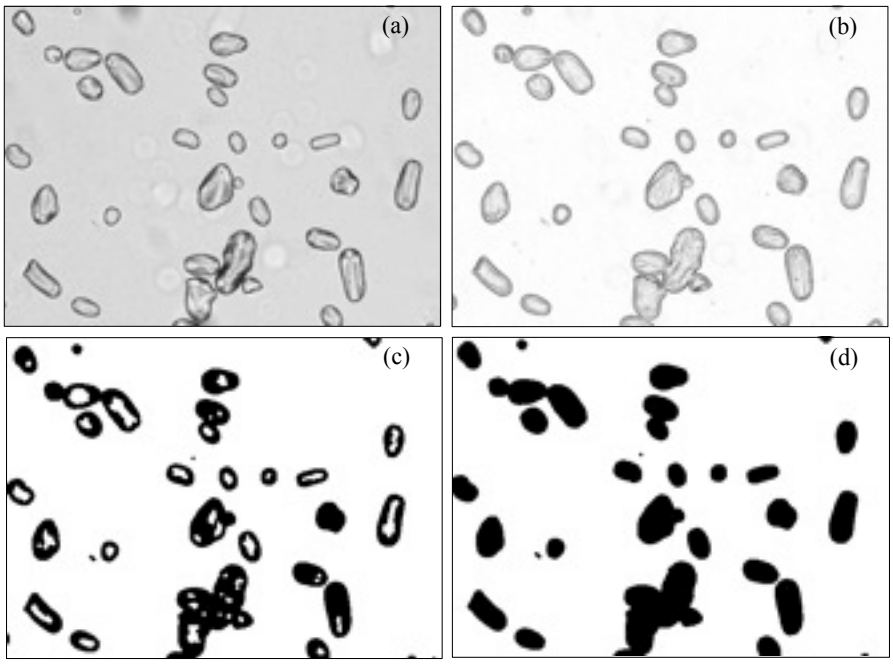

Figure 3: The gradient magnitudes obtained for the image section in (a) are shown in (b). Identification of the shared boundaries from the zero crossings in (c) allows the granules to be separated as shown in (d).

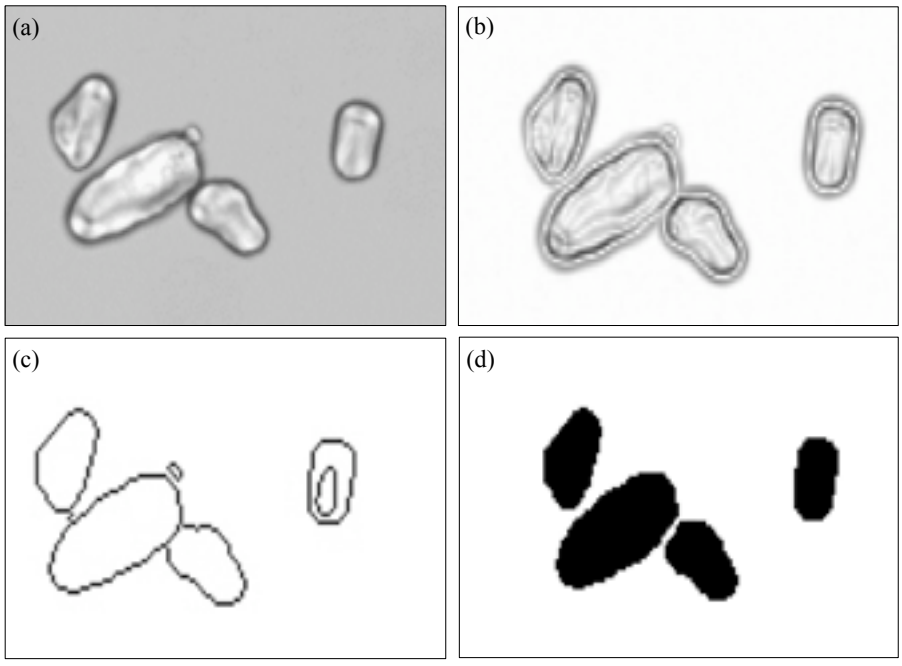

Figure 4: Masked images showing individual potato granules. The white light image is shown in (a) and the corresponding polarized image in (b).

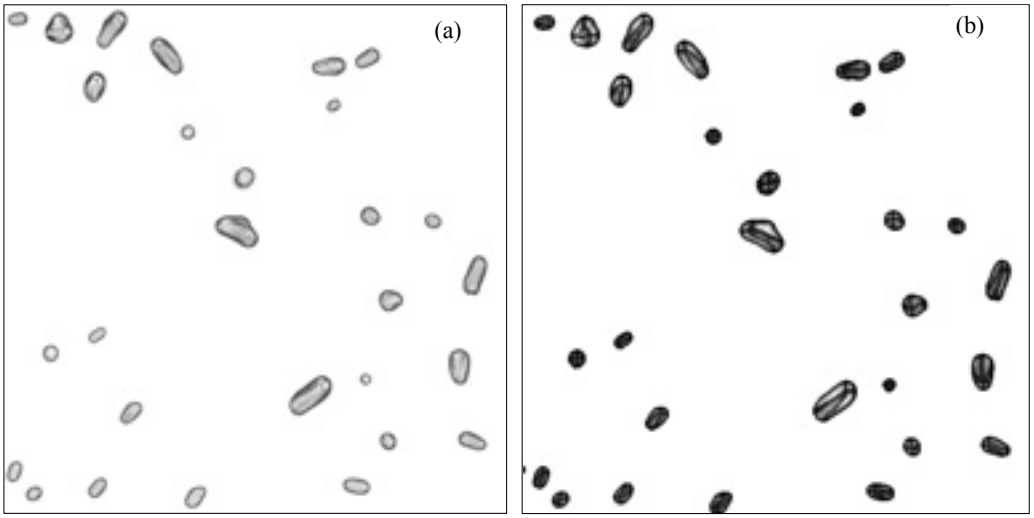

Figure 5: A measure of curvature at the pixel indicated by the grey point can be obtained from the difference in length between the sum of the two dotted lines and the dashed line. In (a) this difference is $\sqrt{5}+2-\sqrt{17}=0.113$, whereas in (b) we have $4\sqrt{2}-4=1.657$, reflecting the greater curvature.

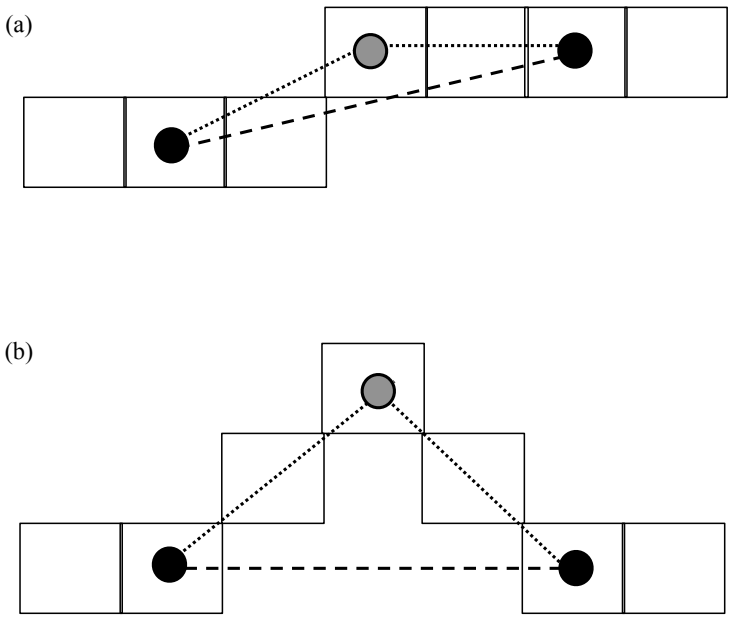

Figure 6: The black line shows the convex hull of the granule, i.e. the smallest convex polygon that contains all the pixels of the granule. The total concavity is defined as the difference between the area of a granule and its convex hull. The maximum concavity is defined as the greatest distance between the boundary of the granule and its convex hull, as indicated by the arrow.

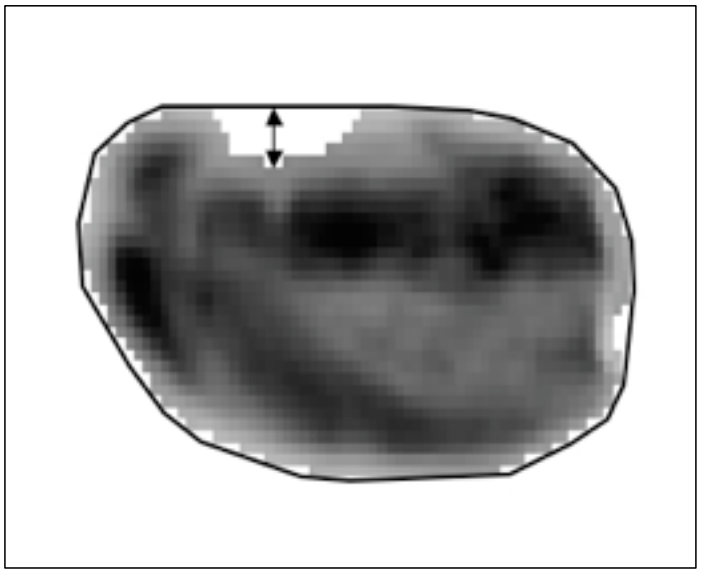

Figure 7: A section of an original polarized image is shown in (a) with the corresponding section shown in (b) after the mask obtained from the white light image has been applied. In (c) the black pixels are those below the threshold, $\mu + 0.1\sigma$, where μ and σ are the mean and standard deviation of the pixels in each masked granule. The percentage of pixels that are black pixels provides a classification variable for each granule. The skeletonized image of the polarization cross is shown in (d).

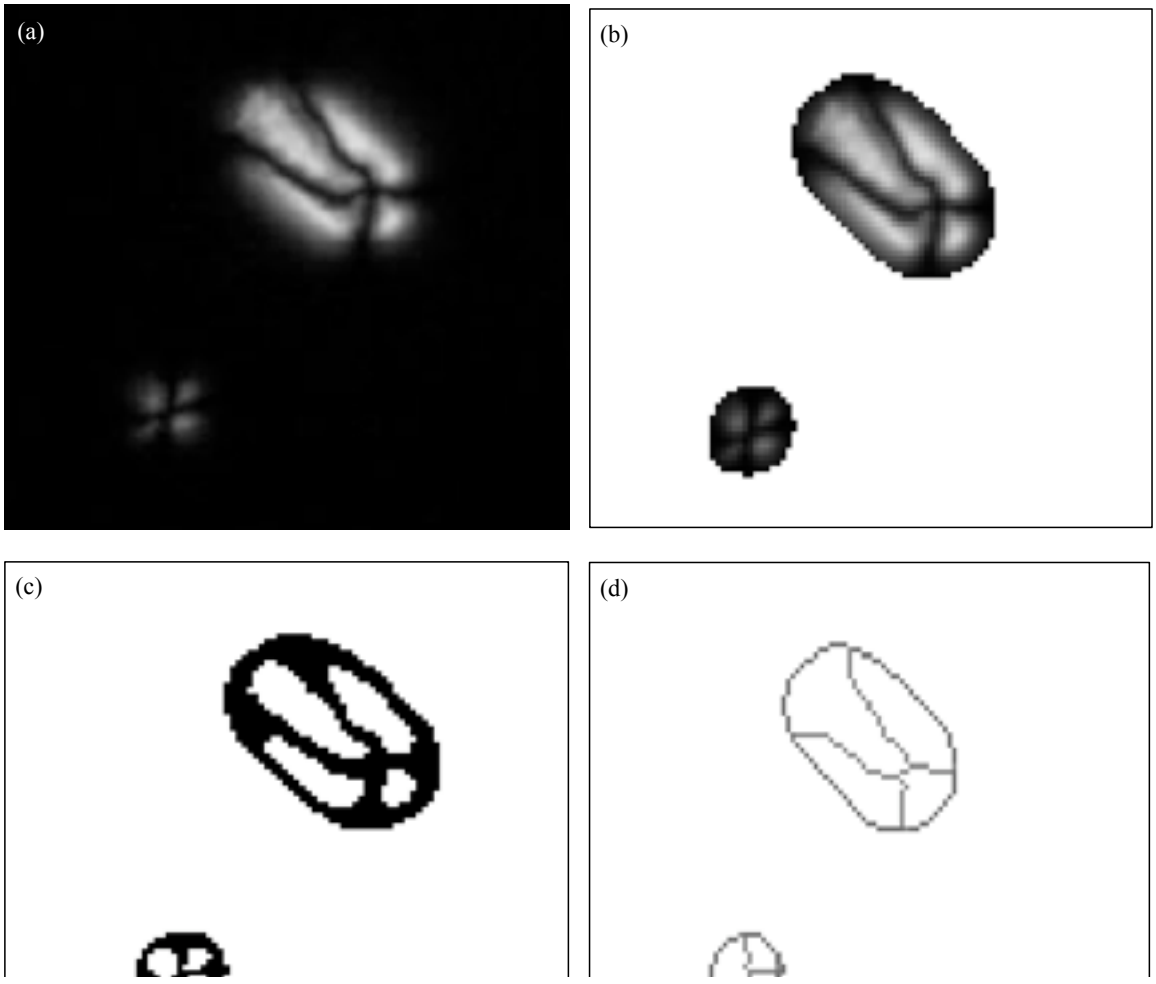

Figure 8: The cracked granules seen in the white light image in (a) and corresponding polarized image in (b) give rise to extra branches in the skeletonized polarisation cross as shown in (c). Extra loops and branches also result from the thresholding procedure. The cross areas obtained for the granules in the polarized image in (d) are shown in (e) with the resulting skeletonization in (f).

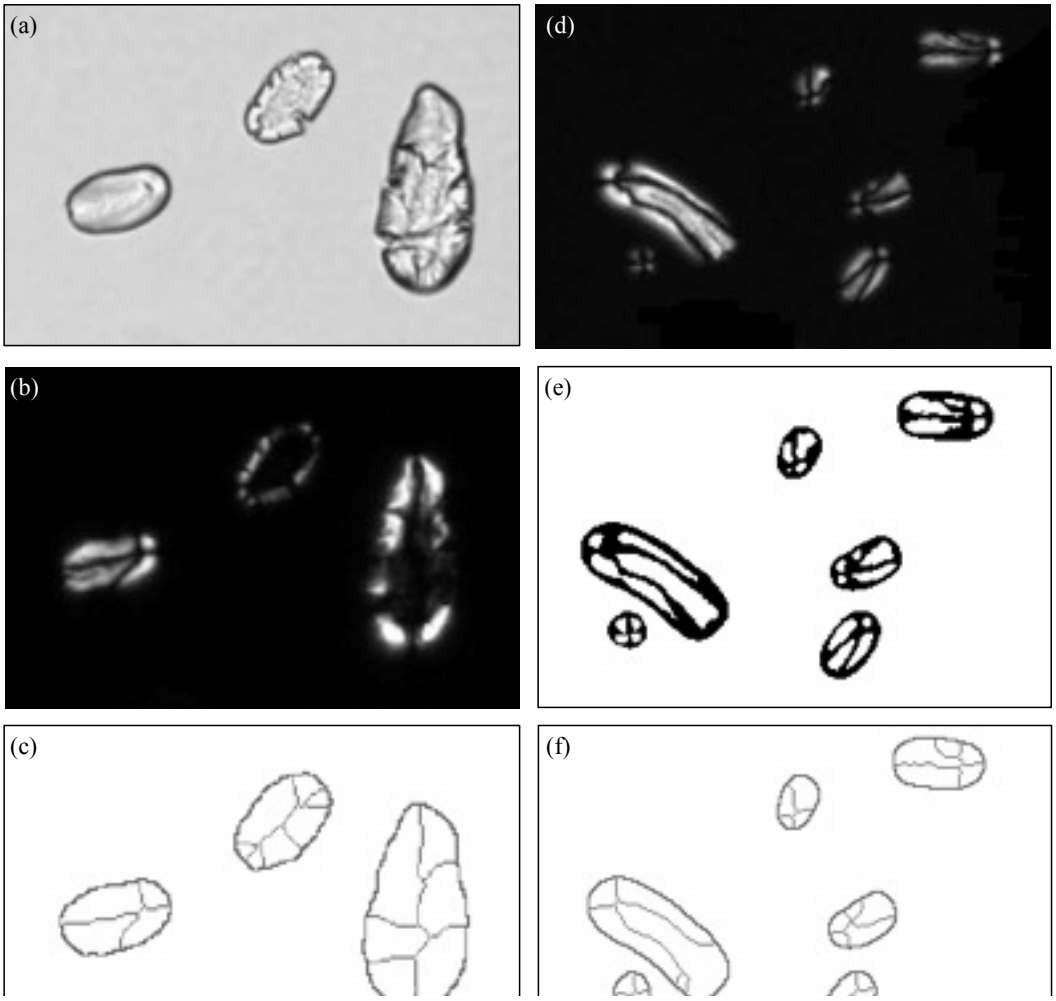

Figure 9: Box-plots showing the distribution of granules sizes for each of the 9 species of granule. Here the size of a granule is determined its length.

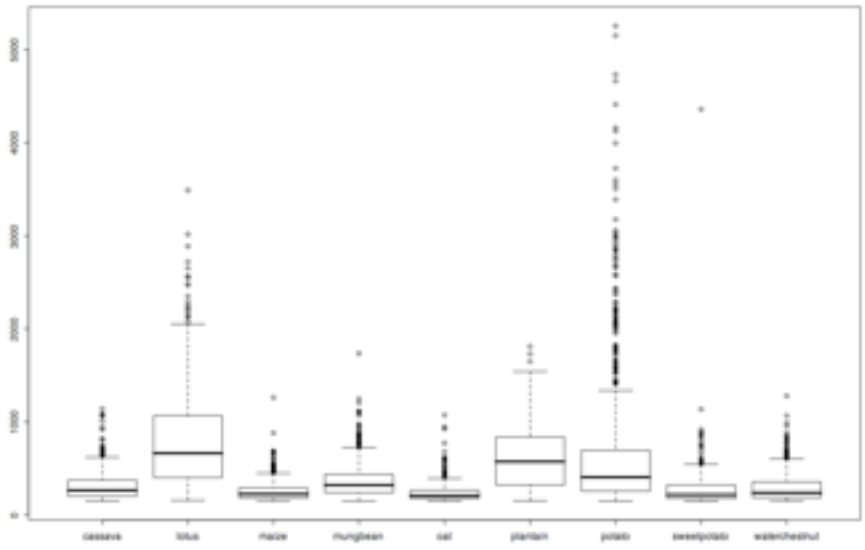

Figure 10: Population pyramid showing the distribution of correctly classified (type 1) and incorrectly classified (type 0) granules according to size.

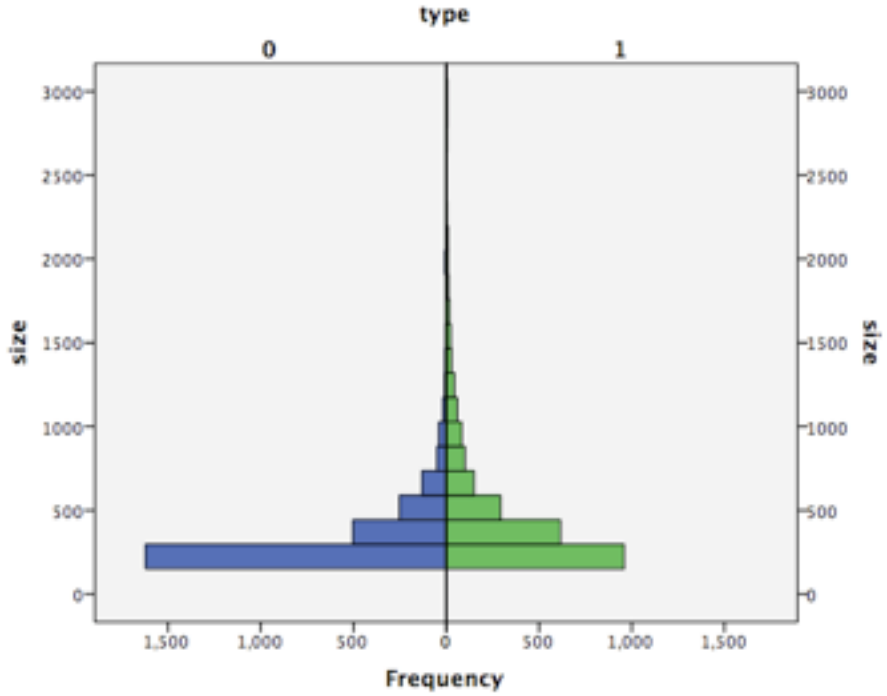

Figure 11. The white light and polarized images for the chimp slide are shown in (a) and (b) respectively. The image obtained by combining these two images in shown in (c) with the resulting masked granules in (d). Each image section shown here is approximately one quarter of the original image.